Introduction à Galaxy

Olivier Inizan
Alban Lermine
Ecole Bioinformatique de Roscoff
18 novembre 2013

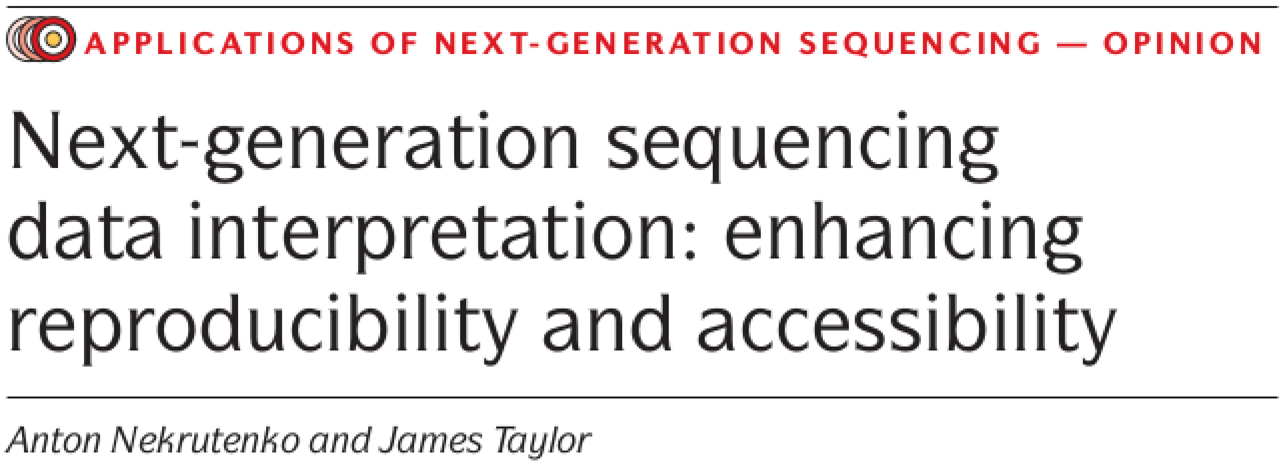
Introduction

- « Big data » problem : a small facet of a much bigger challenge
- Meaningful **interpretation** of sequencing data has become particularly important
- Big data intrepretation constrains
- Galaxy Project : « **democratization** of biomedical computation so that even the smallest research units with modest budgets are capable of carrying out analyses using appropriate tools in a reproducible fashion »

Democratization

- developing **best practices**
- removing obstacles associated with using heterogenous software on complex high performance computing infrastructure : **accessibility**
- promoting the concept of **transparency** and **reproductibility**

Best Practices : emergency !



APPLICATIONS OF NEXT-GENERATION SEQUENCING — OPINION

Next-generation sequencing data interpretation: enhancing reproducibility and accessibility

Anton Nekrutenko and James Taylor

- 1000 Genomes Project : a serie of accepted practices for variant discovery
- Galaxy P.I survey (Anton Nekrutenko and James Taylor)
- 2011 : 299 articles that explicitly cite the 1000 genomes project :
  - 10/299 : used tools recommended by the consortium for mapping and variant discovery
  - 4/299 : used the whole workflow

=> The difficulty of reproductibility

Reproductibilty : is it so easy ?

- NGS analysis is constant flux
- Not only ONE best practice
- Apply to non-model organisms
- Researchers choose to use more straightforward approaches
- Best pratices, accessibility, transparency, reproductibility : the solution with **integrative ressources** ?

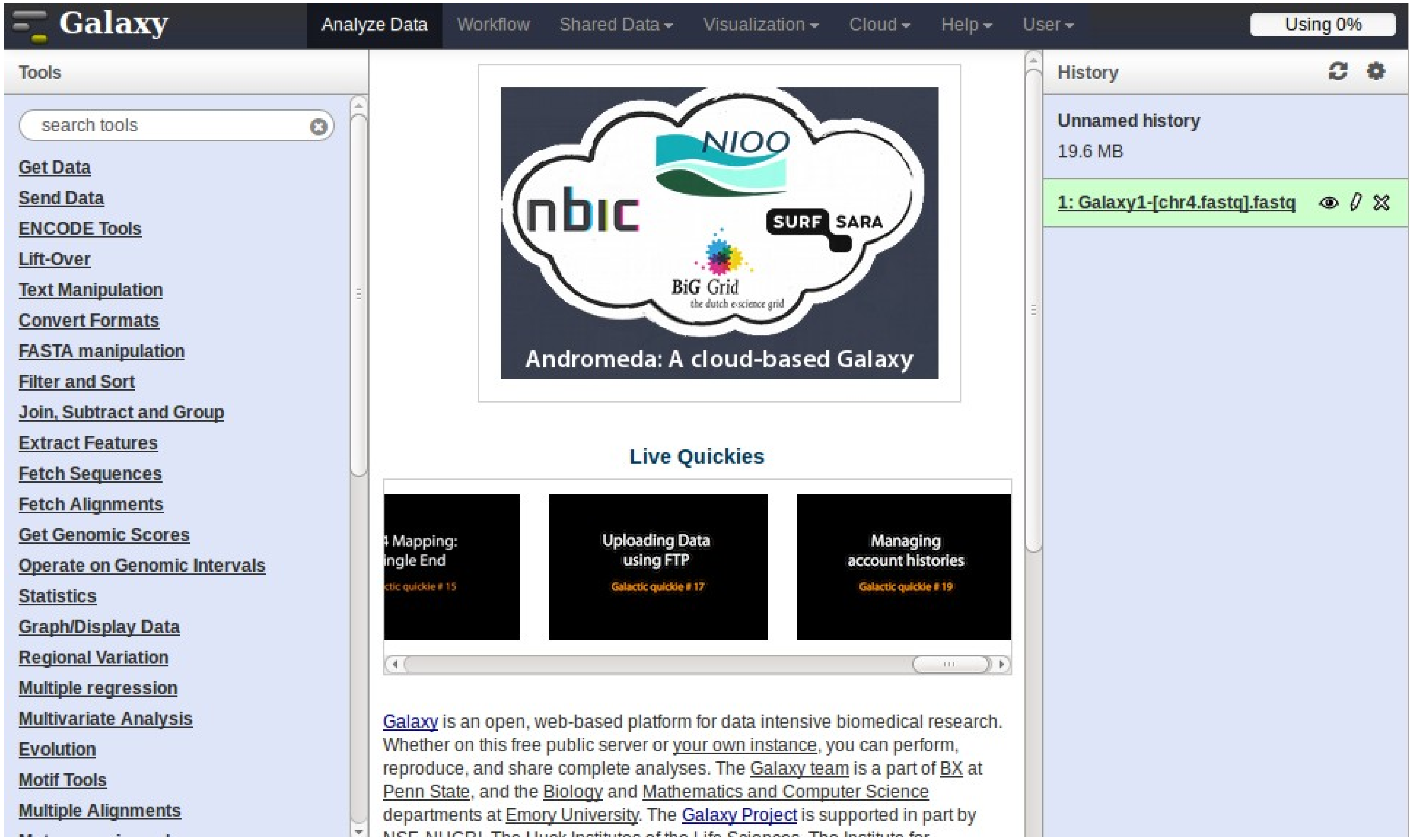
Integrative ressources

- Integratives ressources, integrative frameworks : bring together diverse tools under the umbrella of unified interface
- BioExtract, GenePattern, GeneProf, Mobyle
- Galaxy

Galaxy and « meaningfull interpretation »

- a.k.a how Galaxy embrace accessibility, reproductibility and best practices ?
- **Accessibility** : use computational approaches without programming or informatics expertise
- **Reproductibility** : reproduce experimental results
- **Transparency** : analysis can easily be communicated or understood

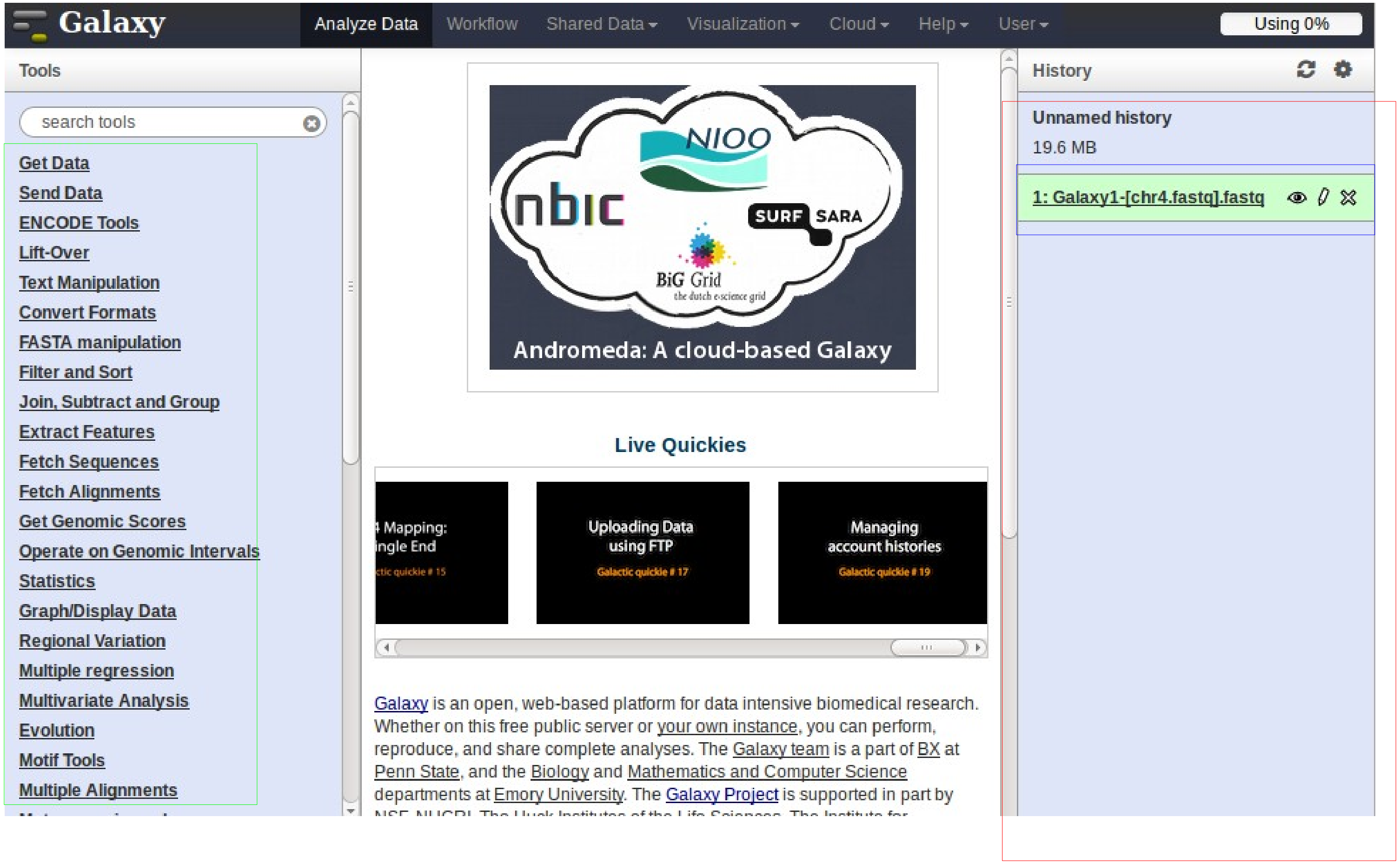
Accessibility



Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for

Provide a unified, web based interface for bioinformatics analysis

Galaxy Items (1 /2)



Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for

tools

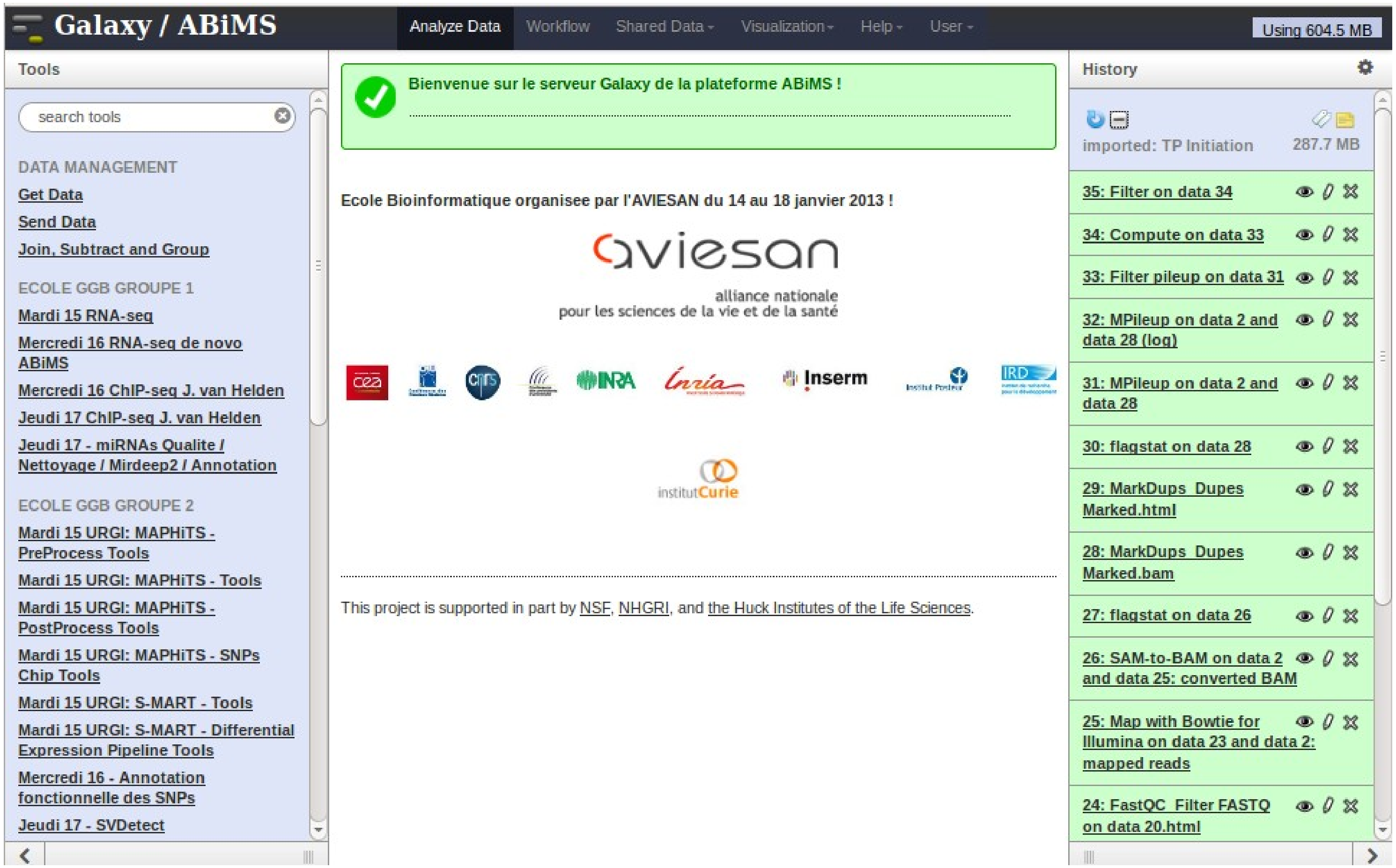
Dataset

history

2 distributions

- 2 distribituions : central (https://main.g2.bx.psu.edu/) and « dist »
- Dist : create your own analysis environment
  - Follow the model Galaxy use for integrating tools
  - A tool = a simple piece of software (cmd line)
  - A developper write a config file (how to run the tool, input and output param)
  - And … Galaxy works with the tool abstractly : automatic generating web interfaces

Your own analysis env, example



DATA MANAGEMENT

Get Data

Send Data

Join, Subtract and Group

ECOLE GGB GROUPE 1

Mardi 15 RNA-seq

Mercredi 16 RNA-seq de novo ABiMS

Mercredi 16 ChIP-seq J. van Helden

Jeudi 17 ChIP-seq J. van Helden

Jeudi 17 - miRNAs Qualite / Nettoyage / Mirdeep2 / Annotation

ECOLE GGB GROUPE 2

Mardi 15 URGI: MAPHiTS - PreProcess Tools

Mardi 15 URGI: MAPHiTS - Tools

Mardi 15 URGI: MAPHiTS - PostProcess Tools

Mardi 15 URGI: MAPHiTS - SNPs Chip Tools

Mardi 15 URGI: S-MART - Tools

Mardi 15 URGI: S-MART - Differential Expression Pipeline Tools

Mercredi 16 - Annotation fonctionnelle des SNPs

Jeudi 17 - SVDetect

Bienvenue sur le serveur Galaxy de la plateforme ABiMS !

Ecole Bioinformatique organisee par l'AVIESAN du 14 au 18 janvier 2013 !

aviesan
alliance nationale
pour les sciences de la vie et de la santé

cea · cnrs · INRA · inria · Inserm · Institut Pasteur · IRD

institut Curie

This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

imported: TP Initiation 287.7 MB

35: Filter on data 34

34: Compute on data 33

33: Filter pileup on data 31

32: MPileup on data 2 and data 28 (log)

31: MPileup on data 2 and data 28

30: flagstat on data 28

29: MarkDups_Dupes Marked.html

28: MarkDups_Dupes Marked.bam

27: flagstat on data 26

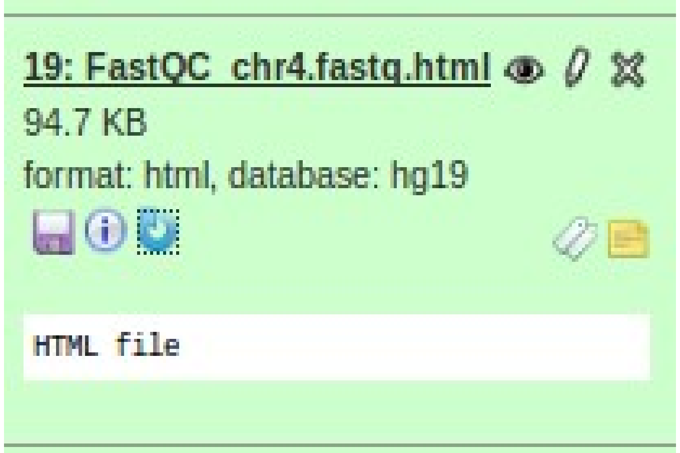
26: SAM-to-BAM on data 2 and data 25: converted BAM

25: Map with Bowtie for Illumina on data 23 and data 2: mapped reads

24: FastQC_Filter FASTQ on data 20.html

Reproductibility

- Galaxy captures **metadata**
- For each step in an analysis : input dataset, tools used, parameters values and output dataset
- With these metadata users can reproduce the analysis



19: FastQC_chr4.fastq.html

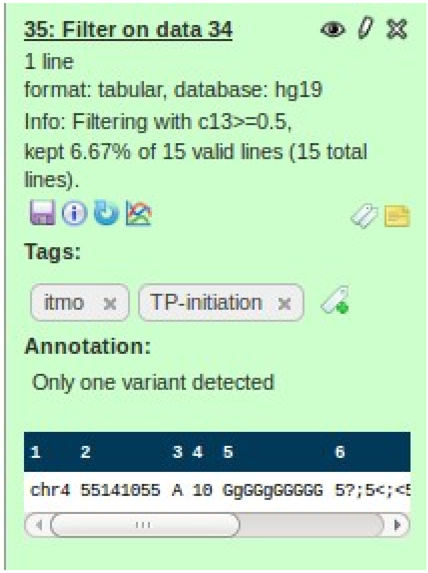
94.7 KB

format: html, database: hg19

HTML file

Reproductibility

- But what about the **intent** of the analysis ?
- Use **annotations** and **tags** (c.f. web practices) to express the intent
- Annotations and tags = user metadata



35: Filter on data 34

1 line

format: tabular, database: hg19

Info: Filtering with c13>=0.5,
kept 6.67% of 15 valid lines (15 total lines).

Tags:

itmo × TP-initiation ×

Annotation:

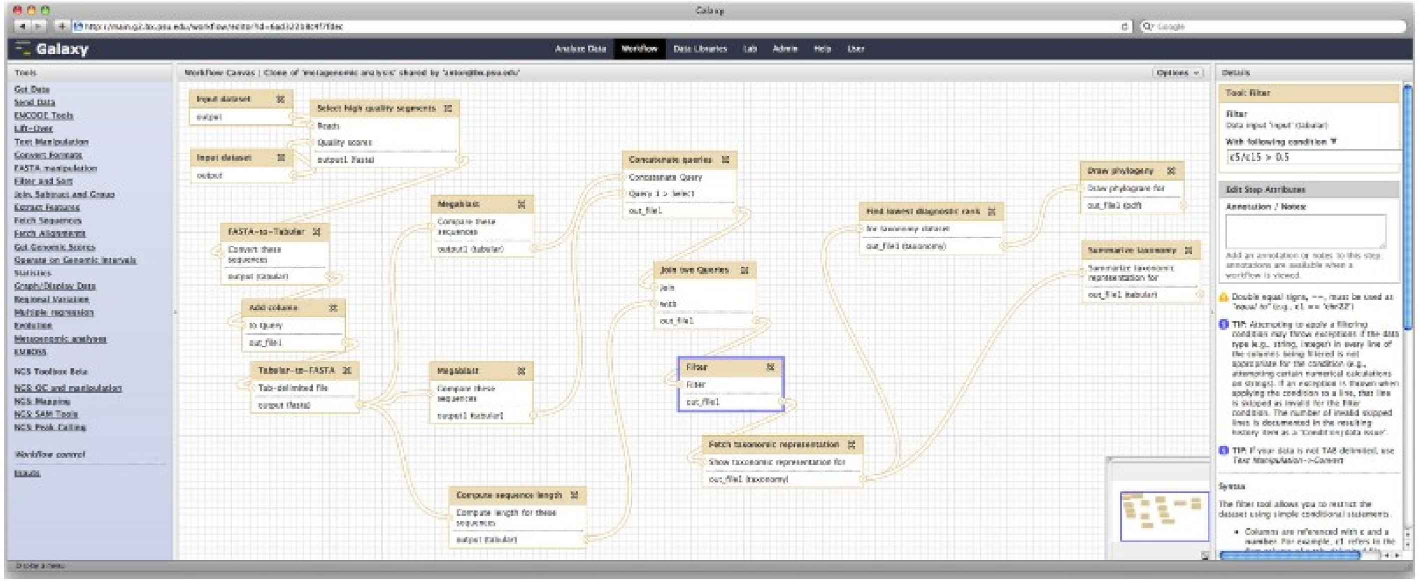
Only one variant detected

1	2	3	4	5	6
chr4	55141055	A	10	GgGGgGGGGG	5?;5<;<

Galaxy Items (2/2)

- And … if I want to reproduce the whole analysis ?
- Galaxy use **workflows**
- Create workflows from scratch, or create from history of your analysis

Workflow (example)



Tool: Filter

Filter

Data input 'input' (tabular)

With following condition: c5/c15 > 0.5

Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

Double equal signs, ==, must be used as "equal to" (e.g., c1 == 'chr22')

TIP: Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

TIP: If your data is not TAB delimited, use Text Manipulation->Convert

Syntax

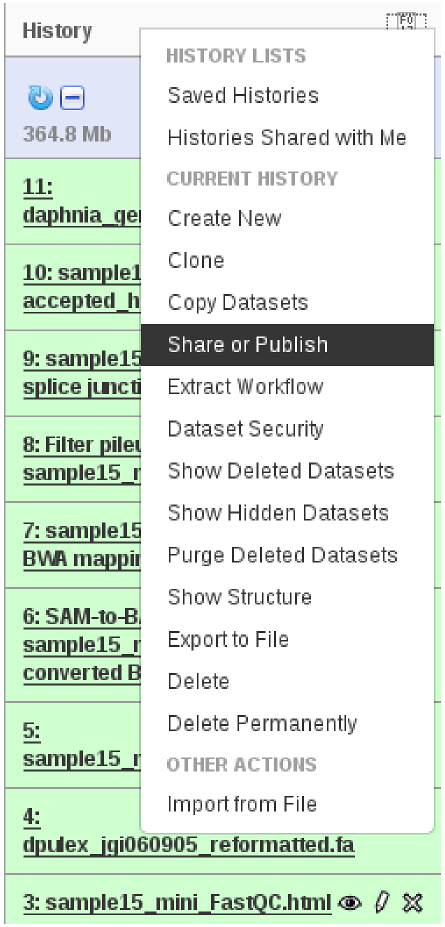
The filter tool allows you to restrict the dataset using simple conditional statements.

- Columns are referenced with c and a number. For example, c1 refers to the

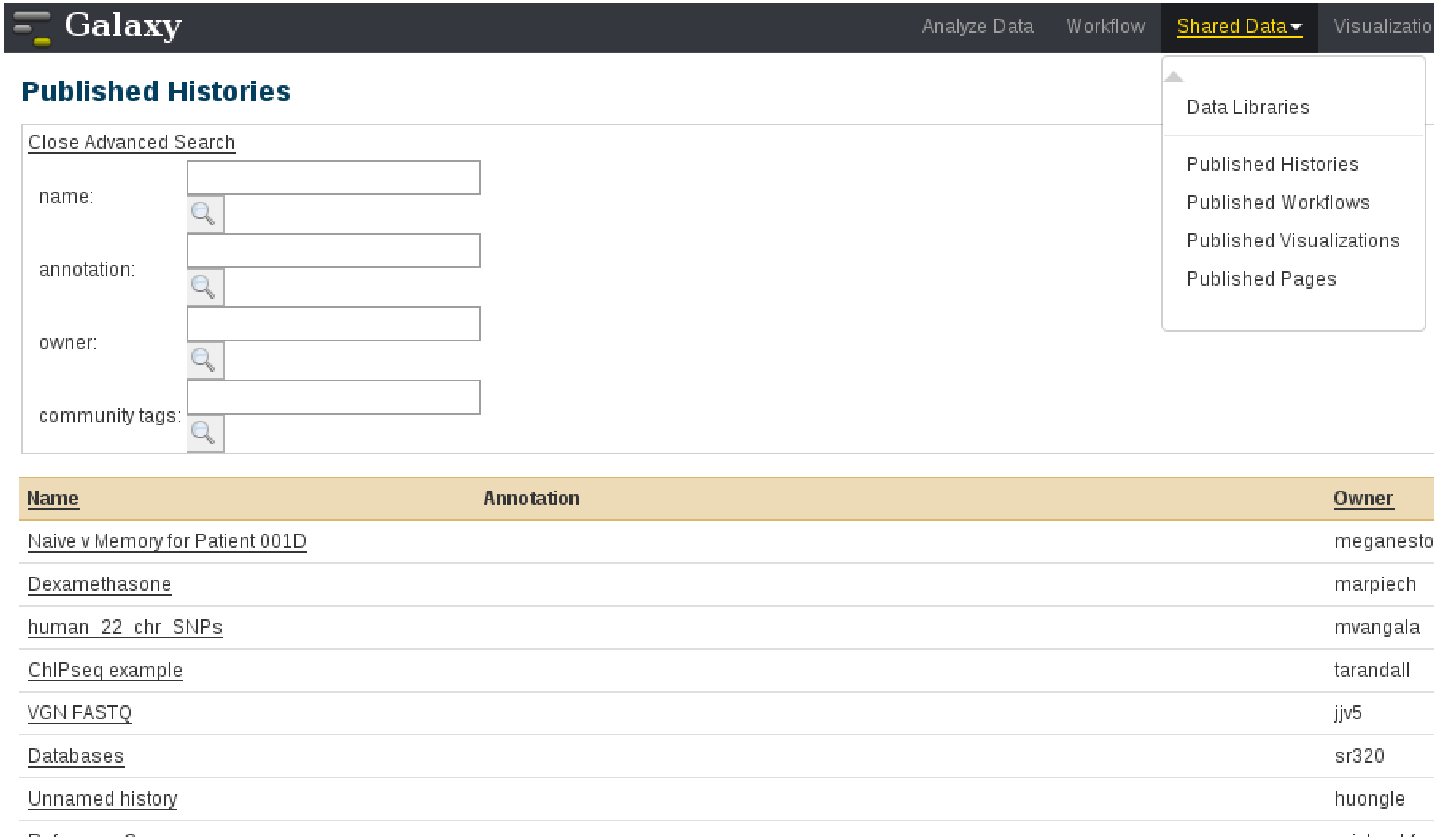
Transparency

- Transparency : enable user to share and communicate their experimental results and output
- **3** elements for Galaxy transparency
- 1 : Galaxy **sharing model** = sharing a Galaxy item* : dataset, histories, visualisation and workflows
- 2 : search shared item from **Galaxy Web Based framework**

Sharing model : example



Search shared item : example



Published Histories

name:

annotation:

owner:

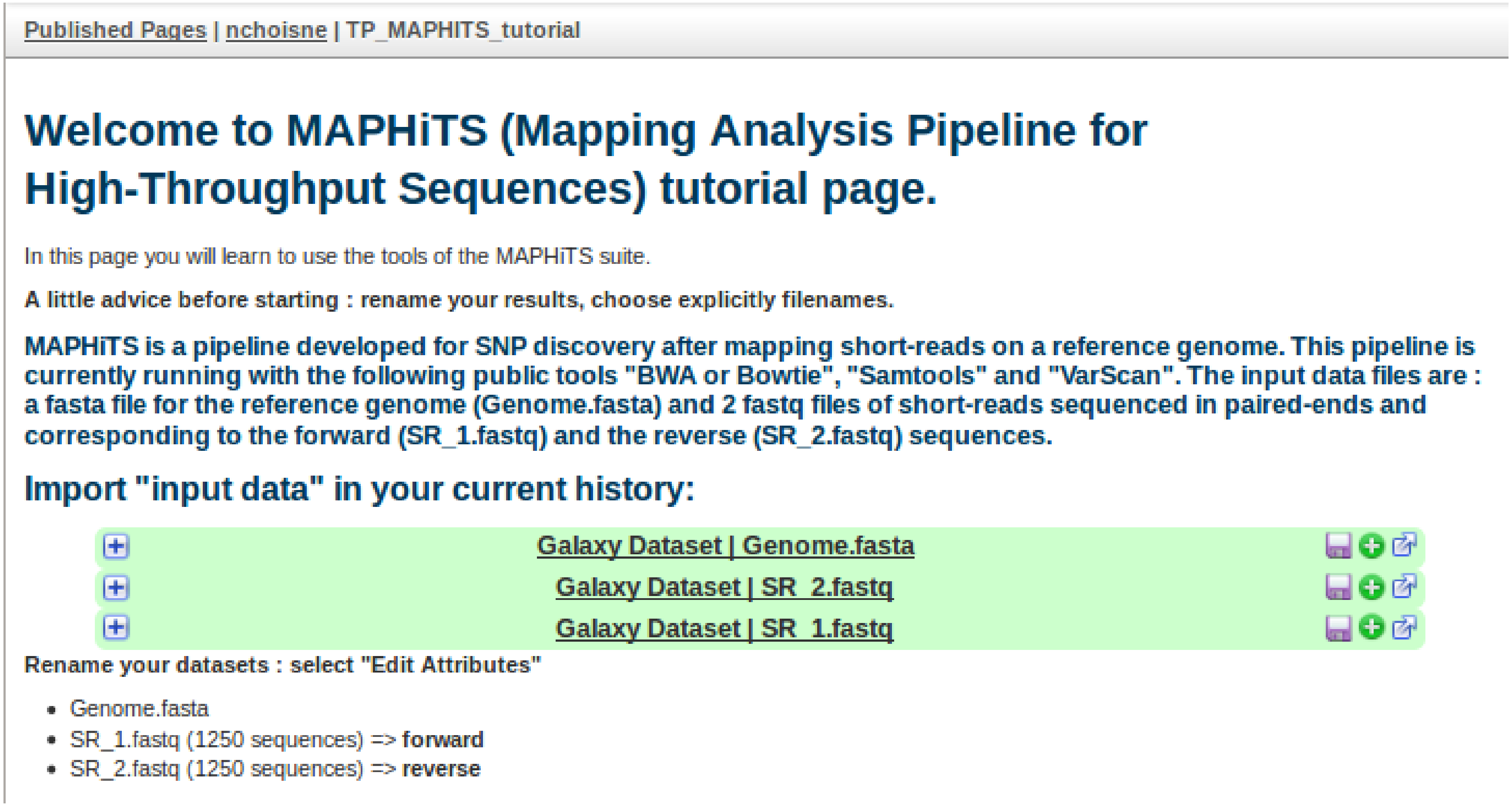
community tags:

Name	Annotation	Owner
Naive v Memory for Patient 001D		meganesto
Dexamethasone		marpiech
human_22_chr_SNPs		mvangala
ChIPseq example		tarandall
VGN FASTQ		jjv5
Databases		sr320
Unnamed history		huongle

Transparency

- 3 : Galaxy **pages**
- Web based document that enable user to communicate their experiement
- A mix of text and graph describing the experiment analysis
- embedded Galaxy items in the page used for the experiment
- Pages and Galaxy sharing model

Page (example)



Published Pages | nchoisne | TP_MAPHITS_tutorial

Welcome to MAPHiTS (Mapping Analysis Pipeline for High-Throughput Sequences) tutorial page.

In this page you will learn to use the tools of the MAPHiTS suite.

A little advice before starting : rename your results, choose explicitly filenames.

MAPHiTS is a pipeline developed for SNP discovery after mapping short-reads on a reference genome. This pipeline is currently running with the following public tools "BWA or Bowtie", "Samtools" and "VarScan". The input data files are : a fasta file for the reference genome (Genome.fasta) and 2 fastq files of short-reads sequenced in paired-ends and corresponding to the forward (SR_1.fastq) and the reverse (SR_2.fastq) sequences.

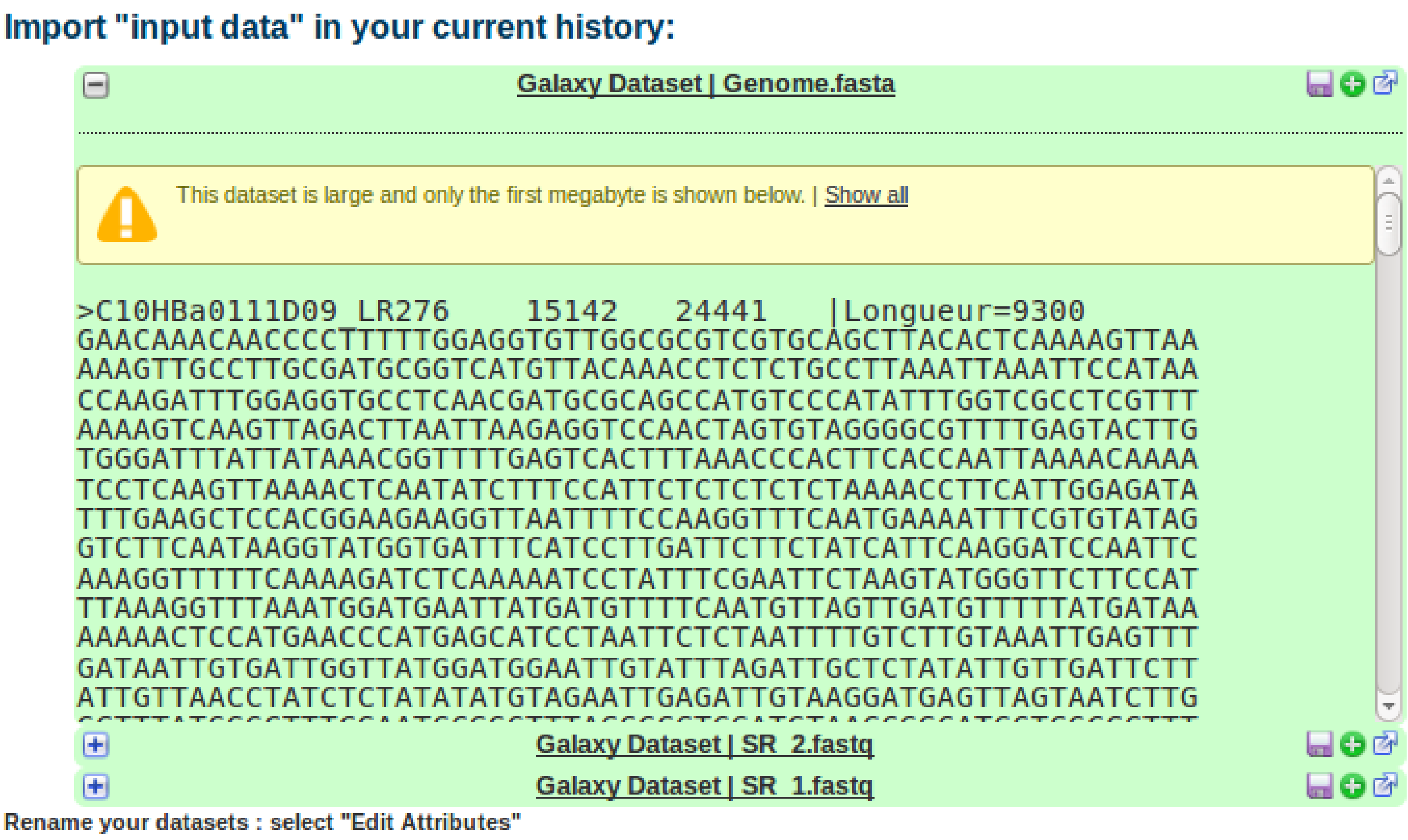
Import "input data" in your current history:

- Galaxy Dataset | Genome.fasta
- Galaxy Dataset | SR_2.fastq
- Galaxy Dataset | SR_1.fastq

Rename your datasets : select "Edit Attributes"

- Genome.fasta
- SR_1.fastq (1250 sequences) => **forward**
- SR_2.fastq (1250 sequences) => **reverse**

Embedded Galaxy item (example)



Import "input data" in your current history:

Galaxy Dataset | Genome.fasta

This dataset is large and only the first megabyte is shown below. | Show all

```
>C10HBa0111D09_LR276    15142   24441   |Longueur=9300
GAACAAACAACCCCTTTTTGGAGGTGTTGGCGCGTCGTGCAGCTTACACTCAAAAGTTAA
AAAGTTGCCTTGCGATGCGGTCATGTTACAAACCTCTCTGCCTTAAATTAAATTCCATAA
CCAAGATTTGGAGGTGCCTCAACGATGCGCAGCCATGTCCCATATTTGGTCGCCTCGTTT
AAAAGTCAAGTTAGACTTAATTAAGAGGTCCAACTAGTGTAGGGGCGTTTTGAGTACTTG
TGGGATTTATTATAAACGGTTTTGAGTCACTTTAAACCCACTTCACCAATTAAAACAAAA
TCCTCAAGTTAAAACTCAATATCTTTCCATTCTCTCTCTCTAAAACCTTCATTGGAGATA
TTTGAAGCTCCACGGAAGAAGGTTAATTTTCCAAGGTTTCAATGAAAATTTCGTGTATAG
GTCTTCAATAAGGTATGGTGATTTCATCCTTGATTCTTCTATCATTCAAGGATCCAATTC
AAAGGTTTTTCAAAAGATCTCAAAAATCCTATTTCGAATTCTAAGTATGGGTTCTTCCAT
TTAAAGGTTTAAATGGATGAATTATGATGTTTTCAATGTTAGTTGATGTTTTTATGATAA
AAAAACTCCATGAACCCATGAGCATCCTAATTCTCTAATTTTGTCTTGTAAATTGAGTTT
GATAATTGTGATTGGTTATGGATGGAATTGTATTTAGATTGCTCTATATTGTTGATTCTT
ATTGTTAACCTATCTCTATATATGTAGAATTGAGATTGTAAGGATGAGTTAGTAATCTTG
```

Galaxy Dataset | SR_2.fastq

Galaxy Dataset | SR_1.fastq

Rename your datasets : select "Edit Attributes"

References and links

- Galaxy Project home page : http://galaxyproject.org/
  - Use galaxy : galaxy-central, a free public server
  - Get a galaxy distribution
  - Learn galaxy : tutorials, screencast
  - Get involved : mailing lists and wiki
- Next-generation sequencing and data interpretation : enhancing reproductibility and accessibility. Anton Nekrutenko ; James Taylor – 2012 – Nature Review Genetics.
- Galaxy : a comprehensive approach for supporting accessible, reproductible and transparent computational research in life science. Jeremy Goecks *et al.* - 2010 – Genome Biology